Nebraska Department of Education

2007 State Writing Assessment: Grades 4, 8, and 11

Standard Setting Study

Final Report

Brett P. Foley, M.S.
Chad W. Buckendahl, Ph.D.

Buros Institute for Assessment Consultation and Outreach

A Division of the Oscar and Luella Buros Center for Testing

University of Nebraska-Lincoln

May 2007

Acknowledgments

We would like to acknowledge our appreciation to several people who assisted us with this Standard Setting Workshop. Sue Anderson, Marilou Jasnoch, and Jackie Naber at the Nebraska Department of Education were very helpful in organizing this workshop. The success of the workshop was due, in large part, to their efforts.

We also want to thank the panelists whose recommendations contributed to the outcome of the standard setting workshop. Panelists who participated in the workshop provided classifications of student performance that resulted in the recommended cut scores. Sue Anderson and Marilou Jasnoch selected the anchor papers and assisted in conducting the standard setting workshop. Without their efforts and diligence, there would have been no study.

Nebraska Department of Education

2007 State Writing Assessment: Grades 4, 8, and 11

Standard Setting Study

Final Report

Introduction

The purpose of this report is to document the procedures and analyses undertaken to recommend performance standards for the Nebraska Department of Education's *State Writing Assessment* administered in grades 4, 8, and 11. The report summarizes the procedures and the results of the standard setting studies and provides recommendations for the establishment of a minimum passing, or cut score for each grade level.

Background

As part of the state assessment and accountability system, Nebraska administers Writing Assessments across the state at selected grade levels (4, 8, and 11). These assessments are used to distinguish between students who have met the state writing standards and those that have not met the state writing standards and may need additional instruction in writing. Because the Writing Assessments are used to classify students in terms of their level of performance in writing, the Department of Education has recognized the importance of using psychometrically accepted methods for setting these performance standards (minimum passing scores).

The writing assessments give students an opportunity to provide a writing sample in response to a narrative (4^{th} grade), descriptive (8^{th} grade), or persuasive (11^{th} grade) prompt. The student writes to the prompt that is provided in a given year. The prompts are scored holistically across six traits on an 10-point scale. Two trained scorers score each paper and the student's score on the paper is the sum of the two scorers' scores. If the two scorers disagree by more than one score point, a third scorer scores the response and an average of the two closest scores is computed.

The purpose of this study was to provide a range of defensible cut scores to the Nebraska Department of Education (NDE) for the *State Writing Assessment* in grades 4, 8, and 11. This report focuses on the results of the standard-setting studies for these three grade levels. The report provides an overview of the methods and procedures for the study. It includes a recommendation for a range of cut scores within which NDE may identify a defensible cut score that will help decide which students in the state have met the writing content standards.

Methods and procedures

Overview of Procedures

Two methods for estimating a cut score were used. Each one relies on different assumptions. The use of these independent methods is intended to provide a more defensible range of possible cut scores, which NDE may use to determine the final cut score. These methods are a) an analytical judgment method and b) a professional judgment method. These methods are described briefly below.

Each of the methods took place on April 19, 2007 in a workshop facilitated by the Buros Center for Testing. The workshop began with an orientation and training activity that included an extended discussion of the test specifications. The training also included a description and discussion of the following student performance levels that were developed by NDE and provided to us for use in the workshop:

1. Beginning: Writing is still under development. Extensive revision or editing would be necessary.
2. Proficient: Writing has more strengths than weaknesses. Some revision or editing would be necessary.
3. Advanced: Writing has many strengths. Only minor revision or editing would be necessary.

The Analytical Judgment Method

One standard setting method used in the standard setting studies is a modification of a method proposed by Hambleton and Plake (2000). This method required panelists to read a set of about 50 papers (described below) and sort the papers into the three broad performance classifications defined above (Beginning, Proficient, or Advanced). After the initial sorting was completed, panelists identified three papers from the "Beginning" papers that were the closest to being in the next higher classification (Proficient). Panelists also identified three papers classified as Proficient that were closest to being Beginning. That is, panelists identified the three best papers in the Beginning classification category and the three worst papers in the Proficient category. Panelists did not know the scores on the papers; instead each paper had an identifying code corresponding to a specific score. The cut score for a panelist was that panelist's mean of the six specific papers that were closest to next higher or lower category. The overall cut score was the average of the individual teacher cut scores.

The about 50 papers were selected using a stratified random sample from the total set of papers. The sampled papers met the following criteria.

1. All score points had at least 2-3 papers with more papers with scores between 2- and 3+ being included.
2. Selected papers were scored correctly and accurately. The basis for scoring was not to be an issue.
3. Selected papers are written legibly and darkly enough that they could be photocopied.

The Professional Judgment method

This method entailed asking panelists to estimate the percentage of tested students in their classes this year who would be classified as Beginning. This was done after all training activities and before participants completed the analytical judgment method. Special forms that also included demographic information to document the level of experience of the panelists were used for this method.

Specific Procedures

The Analytical Judgment Method

The standard setting workshop took place in Lincoln, NE at the Cornhusker Hotel on April 19, 2007. A total of 34 teachers and administrators participated with 12 at 4th grade, 12 at 8th grade, and 10 at 11th grade. All panelists were currently teaching, had recently taught English at their respective grade level, or held positions in their districts related to reading (e.g., Literacy Coach, English Supervisor, Instruction/Literacy Facilitator) and had been exposed to the six-trait writing method used to score the Writing Assessment. Some of the panelists had also participated in the scoring process and/or participated in their respective grade level's writing assessment standard setting previously.

Following introductory comments an orientation and training session was conducted. This session articulated the purpose of the standard setting workshop and detailed the steps to be taken to complete the standard setting process. Training included a discussion of the performance categories (defined above) and a discussion of each of the six traits. Marilou Jasnoch described the six writing traits to the participants. After the large group orientation, the panelists were subdivided into their grade level teams for further training.

In these grade level teams, there was a discussion of the student who was Barely Proficient as the target student. Using performance descriptions derived from the previous standard setting workshops for the state writing assessment, the panelists discussed the skills and performance characteristics of the target student in each of the six traits and holistically. They added to and modified the performance descriptions to better clarify their conception of the Barely Proficient student. These descriptions for each grade level are included as Appendix A. Panelists were advised that they would be reading a large number of papers (50 for grades 4 and 11, 51 for grade 8) and would be making holistic classifications for these papers. These holistic classifications would result in three stacks of papers, those that represented work that was a) Beginning, b) Proficient, and c) Advanced.

Panelists were provided a set of ten papers to practice the process. All panelists received the same papers to rate. These papers were selected such that there were papers that spanned the score range. Panelists made two sorting decisions using these ten papers. First the papers were classified as being Beginning, Proficient or Advanced. After that sorting decision was made, panelists identified the paper from the Beginning papers that was closest to being in the Proficient category. They also selected the paper in the Proficient category that was closest to being in the Beginning category. After these

selections were made, there was a show of hands regarding how each paper was classified. This was followed by a discussion of why panelists made their classification decision.

The training was followed by the professional judgment method and then the operational analytical judgment method. The panelists in each grade level team were provided with copies of 50 papers selected as described above. Panelists made the initial sort into the three broad categories and then selected the three best of the papers classified as being Beginning and the three papers they felt were the worst of those in the Proficient category. Papers were collected and data entered.

Professional Judgment Method

After the practice analytical judgment method was completed, the Professional Judgment method was undertaken. This method entailed having the panelists estimate the percent of students in their classes this year who would be classified as being Beginning.

Results

Analytic Judgment Method

The minimum passing scores are based on the judgments of panelists who made holistic ratings on the 50 (or 51) papers. Each teacher's individual cut score was computed. This involved computing both a mean of the six papers that were just above and just below the performance of the student who was "Barely" Proficient (the target student).

Grade 4

For this grade the recommended cut score using the mean was 3.87. The closest score point to this mean value would be 4.00. The panelists' recommended cut score (4.00), and a range of cut scores plus and minus 1 score point are shown in Table 1. The approximate percent of 4th grade Nebraska students who would be below the cut point is also shown in the Impact column.

For the professional judgment method, panelists' estimated percent of students who will be classified as being Beginning ranged from a low of 1% to a high of 17%, with a mean of 11.14% and a median of 12.00%. The closest score point associated with these impact values is 3.67. It should be noted that these values are based on only the seven panelists currently working with 4th grade students.

Table 1. Analytic Judgment-based cut score and impact and cut scores and impacts within a one score point range for 4th grade.

Range	Cut score	Impact (% below)
1 Score Below	3.67	10.68
Average	**4.00**	**14.27**
1 Score Above	4.33	20.12

Grade 8

For this grade the recommended cut score using the mean was 4.18. The closest score point to this mean value would be 4.33. The mean cut score (4.33), and a range of cut scores plus and minus 1 score point are shown in Table 2. The approximate percent of 8th grade Nebraska students who would be below the cut point is also shown in the Impact column.

For the professional judgment method, panelists' estimated percent of students who will be classified as being Beginning ranged from a low of 0% to a high of 12%, with a mean of 7.00% and a median of 8.00%. The closest score points associated with these impact values is 4.00.

Table 2. Analytic Judgment-based cut score and impact and cut scores and impacts within a one score point range for 8th grade.

Range	Cut score	Impact (% below)
1 Score Below	4.00	7.11
Average	**4.33**	**10.07**
1 Score Above	4.66	12.94

Grade 11

For this grade the recommended cut score using the mean was 3.98. The closest value using the mean would be 4.00. The mean cut score (4.00) and a range of cut scores plus and minus 1 score point is shown in Table 3. The approximate percent of 11th grade Nebraska students who would be below the cut point is also shown in the Impact column.

For the professional judgment method, panelists' estimated percent of students who will be classified as being Beginning ranged from a low of 2.5% to a high of 10%, with a mean of 6.42% and a median of 6.50%. The closest score point associated with the impact values was from 3.67. It should be noted that these values are based on only the six panelists currently working with 11th grade students. Thus, this estimate is likely less stable than the results based on the full panel's judgments on the Analytical Judgment Method.

Table 3. Analytic Judgment-based cut score and impact and cut scores and impacts within a one score point range for 11th grade.

Range	Cut score	Impact (% below)
1 Score Below	3.67	4.76
Average	**4.00**	**6.50**
1 Score Above	4.33	9.23

Evaluation Data

At the conclusion of the workshop, panelists completed an evaluation form consisting of four parts. Part 1 focused on the orientation and training; Part 2 addressed the panelists' levels of comfort and confidence in their Professional Judgment ratings; Part 3 was parallel to Part 2, but focused on the confidence and comfort levels for the Analytical Judgments. Part 4 consisted of closed and open-ended items asking about the

overall success of the workshop and about recommended changes that might be made to improve future workshops. Evaluation comments are shown in Appendix B. Results shown here are for all grade levels. Results were examined by grade level, and were very similar.

Part 1: Training

On a scale ranging from 1 - 6, where 1 = Very Unsuccessful and 6 = Very Successful, all mean ratings fall between 5.5 and 5.6. (Orientation mean =5.6, Training on Analytical Judgments Method mean = 5.5, Description of target students mean = 5.5, Practice with Analytical Judgments Method mean = 5.6, and Overall Training mean = 5.6).

Panelists also rated the adequacy of the time provided for training and orientation. On a six-point scale, where 1 = Totally Inadequate and 6 = Totally Adequate, all mean rating exceeded 5.5. (Orientation mean = 5.6, Training on Analytical Judgments Method mean = 5.7, Description of target student mean = 5.7, Practice with Analytical Judgments Method mean = 5.6, and Overall Training mean = 5.7).

When asked to rate the amount of time allocated to training, the mean rating was 2.1 where a value of 2 was "The right amount of time was allocated to training." Four of 34 panelists felt that too much time was allocated to training; none felt too little time was allocated to training.

Part 2: Professional Judgment Method

The mean panelists' confidence and comfort in making estimate using the Professional Judgment method were 3.7 and 3.8, respectively on a four-point scale (1 = Not Confident/Comfortable and 4 = Confident/Comfortable).

The mean rating for the allocation of time for making the professional judgments was 3.6 on a four point scale (1 = More time needed to be allotted to complete this judgment and 4 = More than enough time was allotted to complete this judgment.); no teacher indicating that the time was insufficient. All panelists felt that there was enough or more than enough time for making these judgments.

Part 3: Analytical Judgments Method

The mean panelists' confidence in classifying papers into three categories was 3.7 on a four-point scale (1 = Not Confident and 4 = Confident). The mean Comfort rating on the same 4-point scale (1= Not Comfortable and 4= Comfortable) for this process was also 3.8.

The final item in Part 3 asked about the adequacy of time allocated for making the analytical judgments. On the four-point scale (1 = More time needed and 4 = More than enough time was allotted), the mean rating was 3.6. None said that more time was needed.

Part 4: Overall

The first item in Part 4 asked about the panelists' confidence in the passing standard that would result from this process. The mean confidence was 3.8 on a four-point scale (1 = Not at all Confident and 4 = Confident). Thus, overall panelists were "Confident" about the appropriateness of the passing standard. None of the panelists had a confidence rating of less than 3.

Two questions asked panelists to rate the success and organization of the workshop (1 = Totally Unsuccessful and 4 = Totally Successful). The mean ratings on these items were both 3.5.

Panelists were given an opportunity to provide comments they felt would be helpful in planning future standard setting studies. Twenty panelists made comments. The comments are attached in Appendix B.

Conclusions and Recommendations

The panelists' recommendations for each grade level are based on considerations of both methods described in the body of this report. For each grade, the cut score based on the Professional Judgment method was lower than the analytical method. However, since several panelists were not currently working with students, we believe that the cut score based on the Analytical Judgment method is more reasonable. For 4th grade, the Analytical Judgment method produced a recommendation for a cut score of 4.00. If a cut score of 4.00 is adopted, approximately 14% of Nebraska's 4th grade students would be identified as having not met the writing standards.

At the 8th grade level, the Analytical Judgment method produced a recommendation for a cut score of 4.33. If a cut score of 4.33 is adopted, approximately 10% of Nebraska's 8th grade students would be identified as having not met the writing standards.

At the 11th grade level, the Analytical Judgment method produced a recommendation for a cut score of 4.00. If a cut score of 4.00 is adopted, approximately 6.5% of Nebraska's 11th grade students would be identified as having not met the writing standards.

References

Plake, B. S., & Hambleton, R. K. (2000). A standard-setting method designed for complex performance assessments: Categorical assignments of student work. Educational Assessment, 6(3), 197-215.

Appendix A

Performance Level Descriptions for Grades 4, 8, and 11

High School (Grade 11): Defining Proficiency for the 6 Traits

Advanced

Stronger	Weaker
Word Choice	
Organization	
Conventions	
Ideas	
Voice	
Sentence Fluency	

Proficient

Stronger	Weaker
Word Choice	Voice
Organization	Sentence Fluency
"Ideas" and "Conventions" - somewhere between strong and weak	

Beginning

Stronger	Weaker
May attempt "Word Choice"	Organization
Voice	Sentence Fluency
	Conventions
	Ideas – sketchy

Barely Proficient 11th grader

Strengths	Weaknesses
Ideas	
Has details	Not fully developed
Commitment to the writer's position	
Generally focused	
Organization	
Functional paragraphing structure	
Sequencing logical	Not always complete
Evidence of Beginning, Middle, and End	Transitions not always there
Voice	
Some conviction	Sometimes forced or mechanical
Appropriate tone for the audience	Inappropriate tone for audience
Word Choice	
Clear and somewhat persuasive – appropriate for 11th grade	Some trite, non-specific language
Details are attempted	Not fully developed or used appropriately in context
Sentence Fluency	
Some variety in structure and length	Phrasing may be more mechanical (related to flow)
Some transitions	Needs more development
Mostly flowing	Few awkward sentences
Conventions	
Errors don't detract from readability (some editing)	Few correct uses of stylistic punctuations.
Basic punctuation, usage (Uses ending punctuations -- . ? !)	
Reader can still understand the message	Some errors in spelling and grammatical usage

Grade 8: Defining Proficiency for the 6 Traits

Advanced

Stronger	Weaker
Word Choice	
Organization	
Conventions	
Ideas	
Voice	
Sentence Fluency	

Proficient

Stronger	Weaker
Word Choice	Voice
Organization	Sentence Fluency
Conventions	
"Ideas" - somewhere between strong and weak	

Beginning

Stronger	Weaker
May attempt "Word Choice"	Organization
Voice	Sentence Fluency
	Conventions
	Ideas – sketchy

Barely Proficient 8th grader

Strengths	Weaknesses
Ideas	
Relevant to topic with some details	Not fully developed
	Not always apparent
Clear	
Organization	
Attempts paragraphs	
Sequencing logical	Not always complete
Evidence of Introduction, Body & Conclusion	Transition not always there
Transitions when attempted are predictable	
Voice	
Some sense of personality Some audience consideration	Sometimes forced or mechanical
Word Choice	
Word used correctly	Word choice may not be creative Some trite, non-specific language
Some sensory details are apparent Some specific words	Not fully developed or used appropriately in context
Sentence Fluency	
Some variety in structure and length	Phrasing may be more mechanical (related to flow)
	Needs more development
Mostly flowing	
Conventions	
Errors don't detract from readability (some editing)	Few attempts to use stylistic punctuations.
Basic capitalization and end punctuation.	
Reader can still understand the message	

Grade 4: Defining Proficiency for the 6 Traits

Advanced

Stronger	Weaker
Word Choice	
Organization	
Conventions	
Ideas	
Voice	
Sentence Fluency	

Proficient

Stronger	Weaker
Word Choice	Voice
Organization	Sentence Fluency
Conventions	
"Ideas" - somewhere between strong and weak	

Beginning

Stronger	Weaker
May attempt "Word Choice"	Organization
Voice	Sentence Fluency
	Conventions
	Ideas – sketchy

Barely Proficient 4th grader

Strengths	Weaknesses
Ideas	
Has some details	Not fully developed
May attempt creativity	Not always apparent
Clear	Multiple off topic details
Organization	
Sequencing logical	Not always complete
Evidence of Beginning, Middle, and End	Transition not always there, not always logical
Hook & Conclusion are attempted	
Voice	
Some personality, evokes some feeling	Sometimes forced Sometimes mechanical
Word Choice	
Clear and somewhat descriptive – appropriate for 4th grade	Some trite, non-specific language
Sensory details may be attempted	Not fully developed or used appropriately in context Unnatural, exaggerated choice of words
Sentence Fluency	
Some variety in structure and length	Phrasing may be more mechanical (related to flow)
Some transitions even if basic	Needs more development
Mostly flowing	
Conventions	
Errors don't detract from readability (some editing)	Few attempts to use stylistic punctuations.
Basic punctuation, usage (Uses ending punctuations -- . ? !)	
Reader can still understand the message. Attempt at paraphrasing.	No attempts at paraphrasing

Appendix B. Comments from Standard Setting Workshop Evaluation

Grade 4

- Hotel accommodations were messed up.
- Program was well planned. I felt comfortable sharing - can't think of any improvements at this time.
- My only question is about the anchor papers. When the raters scored the papers, it was based on the anchors. I wonder if we too should have seen the same anchors, so we were not overly influenced by the writing (high or low) that is currently happening in our classroom. I was thinking of the importance of heading all of us toward the same overall target. Just a question!
- This is very difficult work! I appreciate the opportunity to be a part of the process.
- Sometimes teachers tend to like to "hear themselves talk". I would spend a little less time on discussion. It seemed redundant. I'd rather spend more time on the actual reading…less talking about it.
- As always, the opportunity to talk with others about scoring is great!
- Room reservations - as you are already aware.

Grade 8

- Hotel reservations were a problem.
- We need more space to work and keep organized.
- The room was too crowded and the lighting was poor. We need room to spread out and sort the papers without being elbow to elbow. The staff was pleasant and helpful.
- The rooms were small and did not allow enough space to work. Lighting was dim. Arriving at the Cornhusker and not having a room at 10:00 pm was frustrating. Plus, then having to find a hotel with a room became a challenge. Parking was somewhat of an issue since I was not already at the Cornhusker the night before. In years past it has gone well. This seems to be an isolated incident.
- *Problem with the hotel arrangements. * I feel the NDE "dropped the ball" (hotel sending us out for lunch, crowded rooms, etc.). * A very small working space. * Introductions with teachers would have been beneficial - it is a professional development activity.
- Arrange for larger rooms with table space. - Arrange for lunch on-site. - Provide more info in advance about rooms, parking, etc. I spent 1 1/2 hours driving around before getting a motel room Wednesday night and 1/2 hour parking on the morning of the workshop.
- Could use more space to spread out…
- Parking was difficult.

Grade 11

- I always enjoy working with Buros. Yay Chad B.!!
- Logistics - hotel room "troubles", parking "troubles", Grade 4 not breaking for lunch at the agreed upon time, small spaces in which to work. Orientation presentation - maybe a little more practice so that it will be polished.
- Those of us who have participated in the Buros event before shouldn't have to be here for the orientation. I've done this before - I have taught 6 tracts for many years - I don't need a review.
- More space is needed. We need room to spread out - individual tables would help with the distraction of people in the near proximity. The snafu with the rooms was irritating but not unmanageable. The leadership team was pleasant and helpful. Thanks!
- More work space please. Otherwise, excellent. Thanks.